Laboration 0

Piere Ventura Cruz och Simon Jorstedt

2021-09-26

Data

I uppgift 1 används fyra datafiler bestående av tillverkade observationer hämtade ur ”Graphs in statistical analysis” (The American Statistician, Vol 27, 1973, sid 17-21). I uppgift 2 används en datafil över fyra egenskaper uppmätta för cigaretter från 25 olika märken.

```
data1 <- read.table("ansco1-win.dat", header = TRUE)
data2 <- read.table("ansco2-win.dat", header = TRUE)
data3 <- read.table("ansco3-win.dat", header = TRUE)
data4 <- read.table("ansco4-win.dat", header = TRUE)

data_cig <- read.table("cigarett-win.dat",
                col.names=c("brand","tar","nico","weight","CO"),
                header=FALSE)
data_cig
```

```
##                brand  tar nico weight   CO
## 1             Alpine 14.1 0.86 0.9853 13.6
## 2     Benson&Hedges 16.0 1.06 1.0938 16.6
## 3        BullDurham 29.8 2.03 1.1650 23.5
## 4       CamelLights  8.0 0.67 0.9280 10.2
## 5           Carlton  4.1 0.40 0.9462  5.4
## 6      Chesterfield 15.0 1.04 0.8885 15.0
## 7      GoldenLights  8.8 0.76 1.0267  9.0
## 8              Kent 12.4 0.95 0.9225 12.3
## 9              Kool 16.6 1.12 0.9372 16.3
## 10              L&M 14.9 1.02 0.8858 15.4
## 11       LarkLights 13.7 1.01 0.9643 13.0
## 12         Marlboro 15.1 0.90 0.9316 14.4
## 13            Merit  7.8 0.57 0.9705 10.0
## 14      MultiFilter 11.4 0.78 1.1240 10.2
## 15    NewportLights  9.0 0.74 0.8517  9.5
## 16              Now  1.0 0.13 0.7851  1.5
## 17          OldGold 17.0 1.26 0.9186 18.5
## 18    PallMallLight 12.8 1.08 1.0395 12.6
## 19          Raleigh 15.8 0.96 0.9573 17.5
## 20       SalemUltra  4.5 0.42 0.9106  4.9
## 21         Tareyton 14.5 1.01 1.0070 15.9
## 22             True  7.3 0.61 0.9806  8.5
## 23 ViceroyRichLight  8.6 0.69 0.9693 10.6
## 24    VirginiaSlims 15.2 1.02 0.9496 13.9
## 25    WinstonLights 12.0 0.82 1.1184 14.9
```

Uppgift 1

Vi skall studera data genom att anpassa en (enkel) linjär regressionsmodell för vardera datamängd. Detta ger

```
mod1 <- lm(y ~ x, data1)
mod2 <- lm(y ~ x, data2)
mod3 <- lm(y ~ x, data3)
mod4 <- lm(y ~ x, data4)
```

Sammanfattning av modellen för datamängd 1

```
summary(mod1)
```

```
##
## Call:
## lm(formula = y ~ x, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92127 -0.45577 -0.04136  0.70941  1.83882
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0001     1.1247   2.667  0.02573 *
## x             0.5001     0.1179   4.241  0.00217 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6665, Adjusted R-squared:  0.6295
## F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

Sammanfattning av modellen för datamängd 2

```
summary(mod2)
```

```
##
## Call:
## lm(formula = y ~ x, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9009 -0.7609  0.1291  0.9491  1.2691
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.001      1.125   2.667  0.02576 *
```

```
## x              0.500      0.118   4.239  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6662, Adjusted R-squared:  0.6292
## F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179
```

Sammanfattning av modellen för datamängd 3

```
summary(mod3)
```

```
##
## Call:
## lm(formula = y ~ x, data = data3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1586 -0.6146 -0.2303  0.1540  3.2411
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0025     1.1245   2.670  0.02562 *
## x             0.4997     0.1179   4.239  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 9 degrees of freedom
## Multiple R-squared:  0.6663, Adjusted R-squared:  0.6292
## F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176
```

Sammanfattning av modellen för datamängd 4

```
summary(mod4)
```

```
##
## Call:
## lm(formula = y ~ x, data = data4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.751 -0.831  0.000  0.809  1.839
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0017     1.1239   2.671  0.02559 *
## x             0.4999     0.1178   4.243  0.00216 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.236 on 9 degrees of freedom
## Multiple R-squared:  0.6667, Adjusted R-squared:  0.6297
## F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```

Det första vi noterar är att alla modellerna har (ungefär) samma parametervärden. Modellen $y_i \approx 3 + 0.5 \cdot x_i$ beskriver tydligen alla datamängderna. När vi vidare analyserar parametrarnas p-värden så märker vi att de alla är signifikanta på åtminstone 95%-nivån. Modellerna har även ett justerat R^2-värde (≈ 0.63) som inte omedelbart bör konstateras lågt. Även modellernas p-värden (för parametrarna) samt de justerade R^2-värdena är lika. De fyra modellerna är tydligen i princip identiska. Vi går vidare med att visuellt undersöka datamängderna i figur 1.

```
plot(data1, main="Datamängd 1")
abline(mod1$coefficients[1], mod1$coefficients[2])

plot(data2, main="Datamängd 2")
abline(mod2$coefficients[1], mod2$coefficients[2])

plot(data3, main="Datamängd 3")
abline(mod3$coefficients[1], mod3$coefficients[2])

plot(data4, main="Datamängd 4")
abline(mod4$coefficients[1], mod4$coefficients[2])


#Ifall vi vill plotta residualerna. Jag tror inte vi vill det.
#qqnorm(mod1$residuals)
#qqline(mod1$residuals)

#qqnorm(mod2$residuals)
#qqline(mod2$residuals)

#qqnorm(mod3$residuals)
#qqline(mod3$residuals)

#qqnorm(mod4$residuals)
#qqline(mod4$residuals)
```

I datamängd 1 kan vi se ett möjligt linjärt samband, vilket tycks beskrivas bra med den linjära modellen. I datamängd 3 ser vi ett linjärt beteende i datan, bortsett från en outlier.

Modellen verkar dock inte vara en bra anpassning för datamängd 2, som uppvisar ett troligt kvadratiskt beteende. Modellen är överhuvudtaget inte meningsfull för datamängd 4, som flera identiska x-värden samt en outlier.

Det som rekomenderas för vidare analys är att anpassa en kvadratisk modell för datamängd 2. Eventuellt bör en ny linjär modell anpassas för datamängd 3 (utan outliern), och för datamängd 4 är det snarare aktuellt att använda sig av exempelvis en boxplot eller ett histogram för studera endast y-komponenten.

Låt oss även genomföra en enkel linjär regression på datamängd 1, men med ombytta roller. Datamängd 1 (med ombytta axlar) tillsammans med den nya anpassade reglinjen finns i figur 2.

```
mod1_switch = lm(x ~ y, data1)

plot(data1$y, data1$x,
```

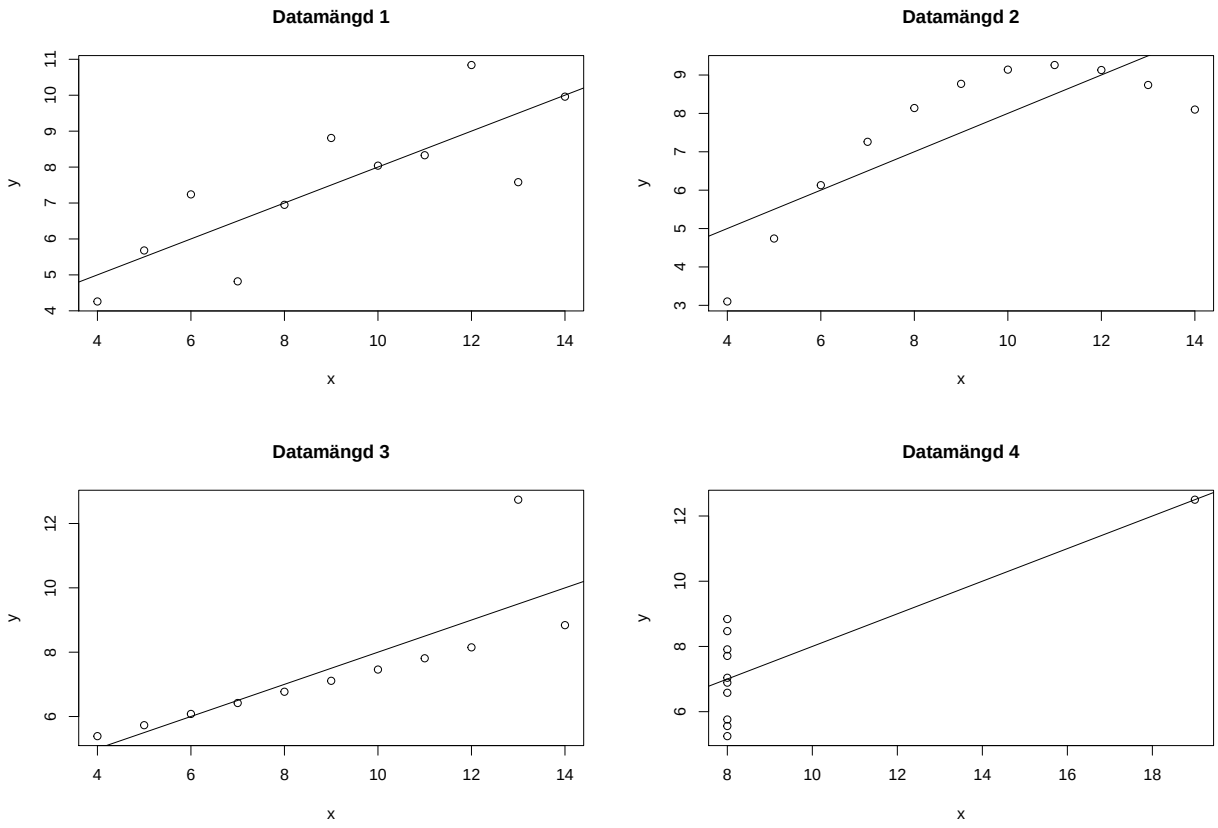

Figure 1: Datamängderna 1-4 tillsammans med anpassade reglinjer.

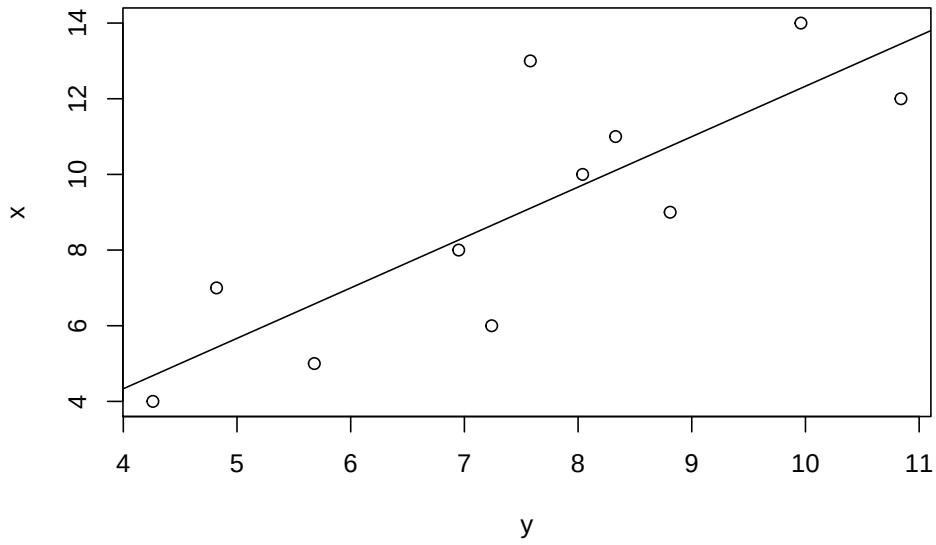

Figure 2: Datamängd 1 samt anpassad regressionslinje med y-komponenten som förklarande variabel.

```
     main="",
     xlab="y",
     ylab="x")
abline(mod1_switch$coefficients[1], mod1_switch$coefficients[2])
```

Sammanfattning av modellen för datamängd 1

Med omvända roller för variablerna.

```
summary(mod1_switch)
```

```
##
## Call:
## lm(formula = x ~ y, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6522 -1.5117 -0.2657  1.2341  3.8946
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.9975     2.4344  -0.410  0.69156
## y             1.3328     0.3142   4.241  0.00217 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.019 on 9 degrees of freedom
## Multiple R-squared:  0.6665, Adjusted R-squared:  0.6295
## F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

Vi har att $y = a + bx \approx 3 + 0.5x$. Med detta i ryggen väntar vi oss inledningsvis att få $x = (y - a)/b \approx 2y - 6$. Men i summary av den anpassade modellen ser vi istället att $x \approx 1.3328y - 0.9975$. Det råder alltså *inte* samma regressionssamband när variablerna är ombytta. Möjligtvis säger detta oss att x-variabeln i datamängd 1 säger mer om y-variabeln, än y-variabeln säger om x-variabeln.

Uppgift 2

Vi skall nu studera hur mängden tjära, nikotin samt vikten för en cigarett påverkar kolmonoxid-utsläppet vid rökning. Inledningsvis anpassar vi en (enkel) linjär modell för vardera tjära, nikotin och vikt som förklarande variabler för kolmonoxidutsläppet (CO). Detta ger följande resultat. Datan samt anpassade (enkla) reglinjer är plottade i figur 3.

```
mod_tar = lm(CO ~ tar, data_cig)
mod_nic = lm(CO ~ nico, data_cig)
mod_wei = lm(CO ~ weight, data_cig)

plot(data_cig$tar, data_cig$CO,
     main="Fig 3.1: CO ~ Tjära",
     xlab="Tjära",
     ylab="CO")
abline(mod_tar$coefficients[1], mod_tar$coefficients[2])

plot(data_cig$nico, data_cig$CO,
     main="Fig 3.2: CO ~ Nikotin",
     xlab="Nikotin",
     ylab="CO")
abline(mod_nic$coefficients[1], mod_nic$coefficients[2])

plot(data_cig$wei, data_cig$CO,
     main="Fig 3.3: CO ~ Vikt",
     xlab="Vikt",
     ylab="CO")
abline(mod_wei$coefficients[1], mod_wei$coefficients[2])

#qqnorm(mod_nic$residuals)
#qqline(mod_nic$residuals)

#qqnorm(mod_tar$residuals)
#qqline(mod_tar$residuals)

#qqnorm(mod_wei$residuals)
#qqline(mod_wei$residuals)
```

Sammanfattning av tjära-modellen

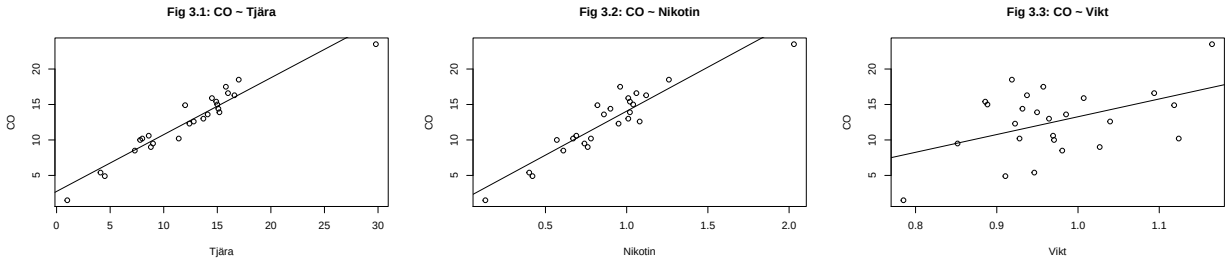

Figure 3: Kolmonoxid plottad mot tjära, nikotin och vikt. Respektive modells reglinje är inkluderad.

```
summary(mod_tar)
```

```
##
## Call:
## lm(formula = CO ~ tar, data = data_cig)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1124 -0.7167 -0.3754  1.0091  2.5450
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.74328    0.67521   4.063 0.000481 ***
## tar          0.80098    0.05032  15.918 6.55e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.397 on 23 degrees of freedom
## Multiple R-squared:  0.9168, Adjusted R-squared:  0.9132
## F-statistic: 253.4 on 1 and 23 DF,  p-value: 6.552e-14
```

Sammafattning av nikotin-modellen

```
summary(mod_nic)
```

```
##
## Call:
## lm(formula = CO ~ nico, data = data_cig)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3273 -1.2228  0.2304  1.2700  3.9357
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.6647     0.9936   1.675    0.107
## nico         12.3954     1.0542  11.759 3.31e-11 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.828 on 23 degrees of freedom
## Multiple R-squared:  0.8574, Adjusted R-squared:  0.8512
## F-statistic: 138.3 on 1 and 23 DF,  p-value: 3.312e-11
```

Sammanfattning av vikt-modellen

```
summary(mod_wei)
```

```
##
## Call:
## lm(formula = CO ~ weight, data = data_cig)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.524 -2.533  0.622  2.842  7.268
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -11.795      9.722  -1.213   0.2373
## weight        25.068      9.980   2.512   0.0195 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.289 on 23 degrees of freedom
## Multiple R-squared:  0.2153, Adjusted R-squared:  0.1811
## F-statistic: 6.309 on 1 and 23 DF,  p-value: 0.01948
```

Sammanfattning av multipel-linjär regressionsmodell

Med tjära, nikotin och vikt som förklaringsvariabler.

```
mod_tnw = lm(CO ~ tar + nico + weight, data_cig)

summary(mod_tnw)
```

```
##
## Call:
## lm(formula = CO ~ tar + nico + weight, data = data_cig)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89261 -0.78269  0.00428  0.92891  2.45082
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2022     3.4618   0.925 0.365464
## tar           0.9626     0.2422   3.974 0.000692 ***
## nico         -2.6317     3.9006  -0.675 0.507234
```

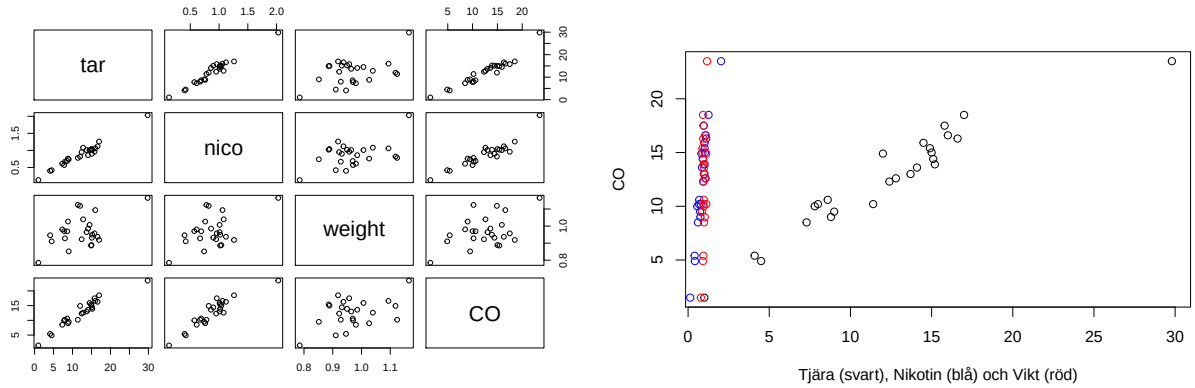

Figure 4: Vänster: Parvisa plottar för variablerna tjära, nikotin och CO. Höger: CO plottad mot tjära, nikotin och vikt.

```
## weight       -0.1305     3.8853  -0.034 0.973527
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.446 on 21 degrees of freedom
## Multiple R-squared:  0.9186, Adjusted R-squared:  0.907
## F-statistic: 78.98 on 3 and 21 DF,  p-value: 1.329e-11
```

Lutningskoefficienterna i de enkla linjära modellerna anger att speciellt nikotin och vikt är associerade med ökat CO-utsläpp. Nikotin-koefficienten har dessutom ett mycket signifikant p-värde ($p \approx 3.31 \cdot 10^{-11}$) samt ett bra R^2 värde ($R^2_{adj} \approx 0.85$). I vikt-modellen har lutningskoefficienten ett signifikant p-värde, men ett mycket lågt R^2-värde ($R^2_{adj} \approx 0.18$) vilket kan tolkas som att vikt-modellen *inte* beskriver variationen i data väl. I tjära-modellen ser vi en positiv association uppbackad av ett mycket signifikant p-värde ($p \approx 6.55 \cdot 10^{-14}$). Även R^2-värdet i tjära-modellen ($R^2_{adj} \approx 0.91$) är högt vilket implicerar att tjära-modellen beskriver variationen i data väl.

I den multipellinjära modellen är lutningskoefficienterna för nikotin och vikt förvånande negativa, med mycket höga p-värden vilket väcker våra misstankar. Tjära-koefficienten är däremot liknande motsvarande i den enkla modellen, återigen med ett lågt p-värde. För att undersöka data vidare plottar vi variablerna parvis i figur 4.

```
pairs(data_cig[, -1])

plot(data_cig$tar, data_cig$CO,
     xlab="Tjära (svart), Nikotin (blå) och Vikt (röd)",
     ylab="CO")
points(data_cig$nico, data_cig$CO, col="blue")
points(data_cig$weight, data_cig$CO, col="red")
```

I figur 4 verkar mängden tjära och nikotin positivt korrelerade. Vikt-datan verkar uppvisa mycket variation, och vara opålitlig för att beskriva CO-utsläppet. Men den parvisa plotten har stora brister; skalorna på x- och y-axlarna maskerar det faktum att nikotin och vikt egentligen är associerade med ökat CO-utsläpp i mycket högre grad än tjära. Till höger i figur 4 framgår det tydligt.

Detta förklarar möjligtvis varför nikotin- och vikt-koefficienterna betedde sig underligt i multipel-modellen ovan. De är mycket starkt korrelerade. Det framgår även en brist i R^2 som förklaringsmått: I de enkla

linjära modellerna ovan hade vikt-modellen ett mycket lägre R^2-värde än tjära-modellen, men i figur 4 (höger) är det tydligt att tjära-datan i själva verket är mycket mer utspridd än vikt-datan.

Vi noterar en outlier i datan (BullDurham). Vi bestämmer oss därför för att genomföra de linjära regressionerna igen, men utan outliern. Först följer de enkla regressionerna, och därefter en multipel-regression med alla tre förklaringsvariabler.

Sammanfattning av tjära-modellen

Exkluderande punkt 3 (BullDurham).

```
mod_t = lm(CO ~ tar, data_cig[-3,])
summary(mod_t)
```

```
##
## Call:
## lm(formula = CO ~ tar, data = data_cig[-3, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7935 -0.7299 -0.3004  1.0733  2.3496
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.41285    0.64822    2.18   0.0403 *
## tar          0.92813    0.05283   17.57 1.96e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.119 on 22 degrees of freedom
## Multiple R-squared:  0.9335, Adjusted R-squared:  0.9304
## F-statistic: 308.6 on 1 and 22 DF,  p-value: 1.964e-14
```

Sammafattning av nikotin-modellen

Exkluderande punkt 3 (BullDurham).

```
mod_n = lm(CO ~ nico, data_cig[-3,])
summary(mod_n)
```

```
##
## Call:
## lm(formula = CO ~ nico, data = data_cig[-3, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2106 -1.1154 -0.1493  1.0656  3.4726
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.2382     1.0827   -0.22    0.828
## nico         14.8600     1.2471   11.92 4.55e-11 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.588 on 22 degrees of freedom
## Multiple R-squared:  0.8658, Adjusted R-squared:  0.8597
## F-statistic:   142 on 1 and 22 DF,  p-value: 4.551e-11
```

Sammanfattning av vikt-modellen

Exkluderande punkt 3 (BullDurham).

```
mod_w = lm(CO ~ weight, data_cig[-3,])
summary(mod_w)
```

```
##
## Call:
## lm(formula = CO ~ weight, data = data_cig[-3, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6387 -2.6256  0.5641  2.8983  7.1507
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.862     10.445   -0.37    0.715
## weight        16.559     10.820    1.53    0.140
##
## Residual standard error: 4.123 on 22 degrees of freedom
## Multiple R-squared:  0.09622,    Adjusted R-squared:  0.05514
## F-statistic: 2.342 on 1 and 22 DF,  p-value: 0.1402
```

Sammanfattning av multipellinjär modell

Med tjära, nikotin och vikt som förklarande variabler. Datapunkt 3 (BullDurham) är utesluten ur data.

```
mod_tn = lm(CO ~ tar + nico + weight, data_cig[-3,])

summary(mod_tn)
```

```
##
## Call:
## lm(formula = CO ~ tar + nico + weight, data = data_cig[-3, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1083 -0.8046 -0.1199  1.0095  2.0501
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5517     2.9713  -0.186 0.854569
## tar           0.8876     0.1955   4.540 0.000199 ***
```

```
## nico          0.5185     3.2523   0.159 0.874941
## weight        2.0793     3.1784   0.654 0.520431
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.16 on 20 degrees of freedom
## Multiple R-squared:  0.935,  Adjusted R-squared:  0.9252
## F-statistic: 95.86 on 3 and 20 DF,  p-value: 4.85e-12
```

I de nya enkla linjära modellerna får vi liknande koefficienter jämfört med då outliern var inkluderad i data. Men vikt-koefficienten har minskat mycket, vilket antyder att outliern hade en stor inverkan på den enkla vikt-modellen. Både tjära och nikotin har fortfarande signifikanta p-värden, medan vikt-koefficienten har ett aningen mindre signifikant p-värde. Den enkla vikt-modellen har ett mycket lågt R^2-värde vilket återigen är anmärkningsvärt eftersom vikt-variabeln tydligt är mindre utspridd än tjära-variabeln som har ett högt R^2-värde.

I den multipel-linjära modellen där outliern är utesluten ser vi återigen det underliga resultatet från den tidigare multipel-linjära modellen, men nu är koefficienterna för nikotin och vikt positiva, vilket är i linje med våra slutsatser från figur 4. Båda koefficienterna är dock låga, och har fortfarande höga p-värden. Multipel-modellen har ett något högre R^2-värde, vilket är väntat eftersom variationen minskar när vi tar bort en variation.

Den nya multipel-modellen (exkluderande outliern) bör presenteras med försiktighet då det är okänt varför outliern stack ut. Det skulle kunna vara ett mätfel, eller så kanske BullDurham tillhör en serie "extrastarka" cigaretter. Utan sådan vetskap bör man därför vara försiktig med antaganden om hur bra modellen beskriver cigaretter som inte ingår i datamängden.

Slutligen kan vi sammanfatta våra resultat med att mängden nikotin och vikten är mycket bra prediktorer för CO-utsläppet. Tjära är också associerat med CO, men i mycket mindre grad. Däremot kan möjligtvis tjära vara relevant som prediktor om man skulle eftertrakta högre precision.

För att konkret kunna mäta och jämföra modellernas prediktiva förmågor skulle vi behöva samla in testdata för att testa modellerna, eftersom hela datamängden har använts som träningsdata. Det är därför svårt och opålitligt att uttala sig om modellerna ovan utan vidare analys.